1. （必填）自己提出的问题的理解（罗列全部）：
2. 提出的问题1：书中提到k-均值算法存在一个空聚类的问题（即在聚类过程中可能有的聚类中心没有被分配任何数据点而使得某些聚类变空，这些聚类被称为空聚类），但是聚类中心本身就是某一个聚类的中心点位置，为什么还会存在空聚类问题呢？

讨论后的理解：在算法的开始，从数据集中随机选取k个数据点作为初始的种子聚类中心，把每个数据点分配给距离它最近的聚类中心。在上述过程中，可能选取的初始聚类中心离所有的数据点都很远，即没有一个数据点被分配到该聚类中心，所以产生了空聚类。

1. 提出的问题2：：在混合属性的处理中，为什么将一个具有多余两个状态的名词性属性或者非对称布尔属性转换为区间度量函数是没有意义的？

讨论后的理解：因为区间度量属性在度量空间相同长度的区间上具有相同的重要性，通过范围标准化方法来标准化区间度量属性时，通过对属性值域的分割将属性值的变化范围转换到0~1之间。对于有两个状态的名词性属性来说，可以将该属性值对应的布尔属性的属性值为1，其余的布尔属性值置为0，即可将该名词性属性转换为区间度量函数。但是如果有多余两个状态的名词性属性，该方法就会失效。对于非对称布尔属性，该属性中的状态具有不同的权重，所以在相同长度的区间上具有不同的重要性，所以没有办法转换为区间度量函数。

1. （必填）别人提出的问题的理解（选择几个问题罗列，并给出理解）：
2. 问题3：使用k-means算法的时候，全局最小值对于大规模数据集来说在计算上是不可行的，那么有没有方法在保证计算相对较小的情况下，获得较高的相对最小值呢？

自己的理解：一种方法是重复计算很多次，直到获得较高的相对最小值；另一种方法是使用二分K-means算法，首先将所有点作为一个簇，将簇一分为二，然后选择能够最大程度降低SSE值的簇继续进行划分，不断重复上述过程，直至划分簇的数目达到要求。

1. 问题4：如何找到最佳的簇数

自己的理解：一是手肘法，核心思想是随着聚类数k的增大，样本划分会更加精细，每个簇的聚合程度会逐渐提高，那么误差平方和SSE自然会逐渐变小。并且，当k小于真实聚类数时，由于k的增大会大幅增加每个簇的聚合程度，故SSE的下降幅度会很大，而当k到达真实聚类数时，再增加k所得到的聚合程度回报会迅速变小，所以SSE的下降幅度会骤减，然后随着k值的继续增大而趋于平缓，也就是说SSE和k的关系图是一个手肘的形状，而这个肘部对应的k值就是数据的真实聚类数。二是轮廓系数法，核心思想为 求出所有样本的轮廓系数后再求平均值就得到了平均轮廓系数。平均轮廓系数的取值范围为[-1,1]，且簇内样本的距离越近，簇间样本距离越远，平均轮廓系数越大，聚类效果越好。那么，很自然地，平均轮廓系数最大的k便是最佳聚类数。

1. （必填）读书计划
2. 本周完成的内容章节：《web data mining》第四章：无监督学习
3. 下周计划：《web data mining》第三章：监督学习